学校编码：**10384**　　分类号_____密级_____
学号：31520101153166　　UDC_____

厦门大学

硕　士　学　位　论　文

基于Viterbi-GMM的文本提示型声纹识别系统的研究及实现

Research and implementation of text-prompted voiceprint recognition system based on Viterbi-GMM

吕伟辰

指导教师姓名：洪青阳　副教授
专　业　名　称：计算机技术
论文提交日期：2013 年　月
论文答辩时间：2013 年　月
学位授予日期：2013 年　月

答辩委员会主席：_____
评　　阅　　人：_____

2013 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

声纹识别技术作为语音识别的一个分支正在蓬勃发展起来，开始应用于各行各业，越来越受到人们的重视。与传统的身份认证技术相比，声纹识别具有数据采集简单，识别方便快捷等优点，是最适合于远程认证的生物识别技术。

目前，声纹识别的方法多种多样，性能也是各有特色。近些年，基于高斯混合模型和通用背景模型(GMM-UBM)的识别方法以其独特的鲁棒性，在众多声纹识别方法中脱颖而出。GMM-UBM 系统既能在文本无关声纹识别中发挥出色，又能在文本相关的识别中表现良好，为不少应用研究者所采用。与此同时，录音回放攻击也一直是声纹识别技术的一大安全隐患，GMM-UBM 系统也不断受到来自录音冒充者的困扰。为此，本文在 GMM-UBM 系统的基础上提出一种文本提示型的方法来克服这一难题。

本文描述的文本提示型系统不同于以往的训练大量文本集的策略，也不是自适应有限声学基元模型的方法，而是通过训练随机文本的语句模型达到近似于文本相关的识别方法。在这个文本提示型的系统中，随机文本的语句模型无法直接获取，本文采用 Viterbi 切分语音的方法，从顺序结构的语音中分离出每个字的语音片段，在语音波形这层将其拼接获得语音句子，将其训练后得到该句子模型，识别时就类似于文本相关的方法。尽管拼接得到的语音句子不够自然，但因文本内容选择的都是汉语数字，语义原本就比较孤立，建立的整体模型并无多少影响，识别性能依然卓越。

在系统性能改进方面，本文从两方面进行着手。首先在 MFCC 特征参数提取上，利用倒谱均值减(CMS)和倒谱方差归一(CVN)的方法，有效消除信道产生的平稳卷积噪音干扰和信道带来的偏移误差，使 MFCC 参数具有更好的鲁棒性。其次，在 GMM-UBM 系统中，考虑到模型是在线训练，需要尽可能加快训练速度，又不失识别性能。本文从影响训练时间的两大因素出发，也就是从高斯混合数的个数以及在训练迭代次数上优化，有效缩短训练时间，获得最优性能。

关键词：GMM-UBM；文本提示型；Viterbi

Abstract

As a branch of speech recognition, voiceprint recognition technology is becoming more and more useful and used in daily life. Compared with the traditional identity authentication technology, voiceprint recognition has the advantages of simple acquisition and convenient identification. What is more, it is the best choice of remote authentication.

There are many methods for voiceprint recognition, and each method has its own characteristic. In recent years, the method of Gaussian mixture model and universal background model (GMM-UBM) has become popular due to its unique performance. It can not only be used for text-independent task, but can be used for text-dependent task. However, the recording playback attack has become a major security risk for voiceprint, and GMM-UBM system is constantly being troubled from the recording impostor. In this paper, we propose a text-prompted system to overcome this problem.

Our method for the text-prompted system is different from the other strategies, which require training a large number of text set. In the text-prompted system, the corresponding model of random text can not be obtained directly, we use Viterbi algorithm to divide the voice of the sequence structure of each word voice clips, and then combine them into voice sentence in the speech waveform layer. This voice sentence will be trained to get the sentence model, which is similar to text-dependent methods. Although the reorganized voice sentence is not so natural, the text content based on ten digits has no much influence on overall model. Our system will thus work well.

To further improve the system performance, we first use cepstral mean subtraction (CMS) and cepstral variance normalization (CVN) method, to effectively eliminate the offset error channel generated by stationary convolution noise interference and channel brings, which make the MFCC parameter more robust. Secondly, we optimize Gaussian's mixed number and the training iterations, to reduce the training time and obtain better performance.

Key words: GMM-UBM; Text-prompted system; Viterbi

目 录

引 言 1

第一章 绪论 2

1.1. 生物识别技术 2

1.2. 声纹识别技术 3

1.2.1. 声纹识别的概念与优势 3

1.2.2. 声纹识别研究历史与现状 3

1.2.3. 声纹识别的分类 4

1.2.4. 声纹识别的常用方法 5

1.3. 课题技术难点 7

1.4. 系统性能评估及工程应用指标 8

1.4.1. 系统性能评估 8

1.4.2. 工程应用指标 9

1.5. 论文的重要工作和安排 10

第二章 语音信号处理 11

2.1. 语音信号的数字模型 11

2.2. 语音信号的预处理 11

2.2.1. 语音信号采样与量化 12

2.2.2. 语音信号的预加重 12

2.2.3. 语音信号的分帧、加窗 12

2.3. 语音信号分析方法的比较 13

2.3.1. 语音信号的时域分析方法 13

2.3.2. 语音信号的频域分析方法 13

2.4. 语音信号的特征参数比较 14

2.4.1. PLP 特征参数 15

2.4.2. LPCC 特征参数 15

2.4.3. Mel 特征参数 15

2.5. 特征参数的改进 17

2.5.1. 倒谱均值减（CMS） 17

2.5.2. 倒谱方差归一（CVN） 17

2.5.3. 一阶差分特征 17

2.6. 语音信号特征选择的依据 18

第三章 文本相关的声纹识别方法 19

3.1. 基于 DTW 文本相关的识别方法 19

3.2 端点检测方法比较 20

3.3 距离测度与失真测度的比较 20
3.4. 特征参数的优化 21
3.4.1. 浮点运算转定点的方法 21
3.5. 语音在模式匹配中存在的问题 22
3.6. DTW 原理 23
3.6.1. DTW 具体解法 24
3.6.2. 文本相关的实验结果 25
第四章 语音切分与重组 28
4.1. 汉语语音特点 28
4.1.1. 声韵母及声调 28
4.1.2. 音节 28
4.2. 语音切分 29
4.2.1. 语音切分的方法比较 29
4.3. 语音合成 35
4.3.1. 语音合成的方法比较 35
4.3.2. 语音合成用于声纹模型的建立 36
第五章 文本提示型系统 38
5.1. 录音回放攻击 38
5.2. 录音冒充解决方法比较 38
5.2.1. 基于信道检测的方法 38
5.2.2. 基于文本提示型的声纹识别的方法 39
5.3. 高斯混合模型（GMM) 39
5.3.1. EM 算法 41
5.3.2. GMM 在声纹确认系统的应用 43
5.3.3. GMM-UBM 声纹模型自适应 44
5.4. 基于 Viterbi-GMM 的文本提示型确认系统 46
第六章 系统实现和实验 49
6.1. 系统设计 49
6.1.1. 硬件平台 49
6.1.2. 软件平台 49
6.1.3. 实验语音库 49
6.2. 文本提示型声纹确认实验 50
6.2.1. Viterbi 切分与手工切分的效果比较 50
6.2.2. 同个人不同数字串之间冒充的识别效果 51
6.2.3. 数据性别上的差异对实验的影响 52

6.2.4. 高斯混合数对系统的影响 53
6.2.5. 系统识别速度的改进 54
第七章 课题总结和展望 56
7.1. 课题总结 56
7.2. 课题展望 56
参考文献 58
致谢 61

Table of Contents

Introduction 1

Chapter 1 prolegomenon 2

1.1. Biometric authentication technology 2

1.2. Voiceprint recognition technology 3

1.2.1. The concept and advantages of voiceprint recognition 3

1.2.2. Research history and current situtation 3

1.2.3. Classification of voiceprint recognition 4

1.2.4. Common methods of voiceprint recognition 5

1.3. Technical difficulties 7

1.4. System performance evaluation and engineering application criteria 8

1.4.1. System performance evaluation 8

1.4.2. Engineering applicationcriteria 9

1.5. Key works and organization 10

Chapter 2 Voice signal processing 11

2.1. The digital model of speech signal 11

2.2. Preprocessing of speech signal 11

2.2.1. Sampling and quantization of speech signal 12

2.2.2. Speech signal pre-emphasis 12

2.2.3. Frames, window of the speech signal 12

2.3. Comparison of speech signal analysis method 13

2.3.1. Time domain analysis of speech signal 13

2.3.2. Frequency analysis method of speech signal 13

2.4. Comparison of feature parameters of speech signal 14

2.4.1. The characteristic parameters of PLP 15

2.4.2. The characteristic parameters of LPCC 15

2.4.3. The characteristic parameters of Mel 15

2.5. Improvement of feature parameters 17

2.5.1. Cepstral mean subtraction (CMS) 17

2.5.2. Cepstral variance normalization (CVN) 17

2.5.3. The first-order differential characteristics 17

2.6. Feature selection for speech signal 18

Chapter 3 The method of text-dependent voiceprint recognition 19

3.1. DTW-based method 19

3.2. Comparison of endpoint detection method 20

3.3. Comparison of distance measure and distortion measure 20

3.4. Optimization of parameters 21

3.4.1. Transformation from floating-point to fixed-point 21

3.5. The pattern matching problem of speech 22

3.6. The principle of DTW 23

3.6.1. The DTW solution 24

3.6.2. Experimental results of the text-dependent 25

Chapter 4 Speech segmentation and reorganization 28
4.1. The characteristics of Chinese speech 28
4.1.1. Acoustic vowel and tone **28**
4.1.2. Syllable **28**
4.2. Speech segmentation 29
4.2.1. Comparative method of speech segmentation **29**
4.3. Speech synthesis 35
4.3.1. Comparison of speech synthesis method **35**
4.3.2. Speech synthesis used to establish the sound-groove model **36**
Chapter 5 The text-dependent system **38**
5.1. Playback attack 38
5.2. Comparison of different methods 38
5.2.1. Method based on channel detection **38**
5.2.2. Method based on the text prompt **39**
5.3. Gauss mixture model（GMM) 39
5.3.1. EM algorithm **42**
5.3.2. The application of GMM in the voiceprint verificaiton system **43**
5.3.3. Voiceprint model adaptation based on GMM-UBM **44**
5.4. Text-prompted voiceprint recognition system based on Viterbi-GMM 47
Chapter 6 System implementation and experiments 49
6.1. System design 49
6.1.1. Hardware platform **49**
6.1.2. Software platform **49**
6.1.3. The speech database **49**
6.2. Experimental results of text-prompted voiceprint recognition **50**
6.2.1. Comparison of Viterbi segmentation and manual segmentation 50
6.2.2. Recognition result of different digit string for the same speaker 51
6.2.3. Influence of gender **52**
6.2.4. Influence of the Gaussian mixture number **53**
6.2.5. Improvement of the system recognition efficiency **54**
Chapter 7 Summary and outlook **56**
7.1. Study summary **56**
7.2. Research prospect **56**
References **58**
Acknowledgement **61**

引 言

随着计算机技术的飞速发展和社会自动化程度的不断提高，人们对方便快捷、更加人性化的人机交互方式的需求日益强烈。一直以来，人们通过键盘、鼠标、显示器等方式与计算机进行互动。最近几年来，通过不断的技术革新，人机交互方式已发生了日新月异的变化，取得了可喜的成果，其中颇具灵气与智能的便是语音交互。可以说，语音交互已经成为了新时代人机交互的关键技术。语音是人类最自然，最习惯的交流方式[1]，与写字，打字的表达方式相比，人们往往更倾向于通过语音来传达自己的思想，如果人与计算机可以通过语音来交流那么不仅充满着趣味性，而且更容易被大众所接受和欢迎。因此，语音声纹识别技术因其实用价值得到了越来越多人的研究，但是一些常见问题也不断出现，如模仿冒充、录音冒充，成为声纹识别技术的安全隐患。

第一章 绪论

1.1.生物识别技术

随着人工智能技术和生物科技的飞速发展，生物识别技术已经得到了广泛的关注。生物识别技术主要是指通过人类自身生物特征进行身份认证的技术，这里的生物特征一般不可复制。生物识别的特征大致包括身体特征和行为特征两类，其中身体特征包括：指纹、静脉、掌形、唇膜、牙齿等，行为特征包括：语音、签名、行走步态等[2]。如今，生物识别技术已经在很多领域有着实际应用，如指纹、人脸等识别技术在安防、考勤等领域广泛应用。

语音是在说话人和听众之间相互沟通的信息，传递的介质是声音波形，不同人之间的说话声音特征会有所区别，这是因为每个人声音的强度和频率组成各不相同，即有不同的音色，因此人们可以通过声音来辨别不同的人。人类声音特征有所区别，本质上是由于不同人之间的发声器官构造差异性所致，计算机如能像人类一样自动辨别出不同人之间的声音特征，那么应用前景将十分广阔。

早在二战期间的美国就已经开始利用声纹来分析识别说话人。1941 年,美国贝尔实验室发明了声谱仪,标志着现代声纹识别技术的开始。二战结束后,贝尔实验室的物理学家 L·G·Kesta 受到美国司法局的委托，利用声谱仪对声纹鉴定进行了系统科学的研究，在研究中发现采用该方法能达到 99.65%的识别率，为美国法院在取证方面做出了极大的贡献。我国对声纹识别技术的研究和应用开始于 20 世纪 80 年代，起步较晚，但发展迅速，且成绩斐然，甚至在某些方面已经处于世界领先水平。